Computational Tools for the Prediction and Classification of Estrogenic Compounds

Project Scope

Growing concern related to endocrine disrupting chemicals (EDCs), exogenous chemicals that affect the endocrine systems of humans and wildlife by mimicking endogenous hormones, has led to federal legislation mandating that the U.S. Environmental Protection Agency (EPA) develop a screening and testing program to assess the EDC activity of these widely prevalent chemicals. This program has the potential to be very labor- and time-intensive, involving *in vitro* and *in vivo* biological testing of many agents against multiple biological endpoints. To maximize efficiency and minimize expense of testing, the use of computer-based tools to enable rapid and large-scale screening of chemicals for their potential EDC activity is essential. In addition, such computational tools can be used to obtain a better understanding of the biochemical mechanisms of EDC toxicity.

The main objectives of this research were to:

- Construct and validate quantitative tools for predicting potential EDC activities of chemicals based on Quantitative Structure-Activity Relationship (QSAR) models based on complimentarity to the chemical structure for two estrogen receptor subtypes (α and β).
- Explore correlations between the ligand-estrogen receptor binding activity predicted by the QSAR models for one species and the experimentally determined binding activities in other species (inter-species extrapolation).

The research was intended to make a major contribution to the development of computational models for predicting endocrine disrupting effects both *in vitro* and *in vivo* for a large number of natural and synthetic chemicals from numerous structural classes. Inappropriate estrogen receptor (ER) activation and/or inhibition by EDCs has been shown to disrupt the transcription of genes necessary for sexual reproduction and differentiation. Computational models offer significant promise as guides for prioritizing EDCs for investigation based on their predicted biological activities. These models also may have significant utility in screening compounds for potential regulatory attention.

Grant Title and Principal Investigator

Computational Tools for the Prediction and Classification of Estrogenic Compounds

William J. Welsh - University of Missouri - St Louis

EPA STAR Grant #R826133

Key Findings and Implications

- A range of CoMFA-based QSAR models were developed that successfully predicted the estrogen receptor binding activity of a wide range of chemicals.
- A computational technique called *Shape Signatures* was developed as a rapid a simple approach to summarizing and comparing the shapes and structural features of molecules to each other and to target receptor sites/subsites.
- The *Shape Signatures* approach will enable rapid shape comparisons of a library of diverse compounds against a target receptor (e.g., estrogen receptors) or against other compounds (e.g., known estrogenic compounds).
- This technology shows promise for the rapid prioritization of potential EDCs for timely and cost-effective biological screening.

Research under this grant resulted in 11 peer-reviewed publications.

Project Period: December 1997 to December 2000

A range of new QSAR methods has recently been developed. QSAR models, such as those obtained using comparative molecular field analysis (CoMFA), have been demonstrated to be useful for risk assessment. Estrogenic EDCs are structurally diverse and a wide range of distinct chemical families have been shown to possess estrogenic activity. Therefore, QSAR models must be capable of relating this structural diversity to biological activity. For the most part, previous models for predicting estrogenic activity have been constructed from data sets with limited structural diversity. In this research, CoMFA models were constructed based on biological data for a structurally diverse set of EDC compounds from a number of chemical families.

QSARs attempt to predict biological activity (for example, receptor binding) based on the physical and chemical characteristics of test chemicals using statistical or pattern recognition methods. The approach used in this research employed:

- A combination of two QSAR paradigms (classical descriptor-based QSAR and 3D-QSAR),
- Three programs for generating more than 400 molecular descriptors necessary for modeling *in vivo* activity, and
- Linear (Partial Least Squares) and non-linear (Artificial Neural Networks) regression models and Genetic Algorithms (GAs) to select an optimal set of molecular descriptors for modeling and predicting estrogenicity.

Relevance to ORD's Multi-Year Research Plan

This project contributes directly to the Long-Term Goal 3 (LTG-3) of ORD's MYP by developing quantitative structure activity relationship (QSAR) models to serve as screening/priority setting tools for EDCs.

Using suite of computational tools developed during this research, a large number of structurally-diverse chemicals can be rapidly prioritized according to their potential to cause endocrine disruption, reducing the need for costly *in vitro* and *in vivo* biological testing. In addition, these tools expand the understanding of the biochemical mechanisms of EDC toxicity. This novel computational technology, highly complementary to QSAR-based approaches, may also be a useful decision support tool for regulators involved in programs to limit harm to humans from EDC exposures.

Project Results and Implications

In vivo and *in vitro* studies suggest that the selective effects of estrogenic compounds may arise in part by the action of different subsets of estrogen-responsive promoters by the two ER subtypes: ER-α and ER-β. The major gene-expression effects of the active form of estrogen, 17b-estradiol, appear to be mediated through these two members of the steroid hormone receptor superfamily. Although the X-ray crystal structure of the ER-α has been elucidated, the three-dimensional structure of ER-β was not publicly available at the time this research was conducted. Based on the significant structural conservation across members of the steroid hormone receptor family and the high sequence homology between ER-α and ER-β, a homology model of the ER-β structure was developed. Using the crystal structure of ER-α and the homology model of ER-β, a strong correlation was demonstrated between computed values of the binding energy and published values of the observed relative binding affinity (RBA) for a variety of compounds for both receptors.

CoMFA was employed to construct a 3D-QSAR model to identify the structural prerequisites for ligand-ER binding and to discriminate ER-α and ER-β in terms of their ligand-binding specificities. The model was developed based on data from a set of 31 structurally diverse compounds for which competitive binding affinities have been measured against both ER-α and ER-β. Structural alignment of the molecules in relation to the ER-α and ER-β binding sites was achieved using the steric and electrostatic alignment (SEAL) algorithm. The final CoMFA models, generated by correlating the calculated 3D steric and electrostatic fields with the experimentally observed binding affinities using partial least-squares (PLS) regression, exhibited excellent self-consistency, as well as high internal predictive ability based on cross validation. CoMFA-predicted values of RBAs for a test set of compounds outside of the training set were also consistent with experimental observations. These resultant models can serve as guides for the rational design of ER ligands that possess preferential binding affinities for either ER-α or ER-β. These models may also prove useful in risk assessment programs to identify suspected EDCs.

3D-QSAR models derived from CoMFA were also constructed using yeast-based reporter gene assay data for a different set of 53 compounds from several chemical classes (e.g., steroids, synthetic estrogens, phytoestrogens, antiestrogens, DDT, polychlorinated biphenyls, other industrial chemicals), that exhibit a high degree of self consistency and predictive ability for estrogenic activity. Three different alignment schemes (SEAL, atom fit, field fit) were tested to obtain the best CoMFA model. The field-fit alignment scheme provided the best predictive model.

Another 3D-QSAR CoMFA model for estrogenic activity was constructed for 53 2-phenylindole compounds. The model was developed using data from an ER binding assay in calf uterine cytosol; measured RBA values spanned a range of four orders of magnitude. The model exhibited excellent self-consistency and predictive ability across this wide range of activity. To examine the possibility of interspecies extrapolation, the calf ER-derived CoMFA model was used to predict the calf ER RBA values for 14 estrogenic compounds for which human ER RBA values were available. The correlation between these predicted calf ER RBA values and corresponding experimental human ER RBA values was high, especially considering the species-to-species differences. A separate QSAR model constructed using classical physicochemical descriptors as the independent variables was shown to be inferior in statistical quality to 3D-QSAR models derived using CoMFA.

Based on the experience accumulated in the development of the models described above, a novel technique was developed for rapidly comparing the shapes of molecules to each other and target receptor sites/subsites. This approach, called *Shape Signatures*, involves generating compact representations of shape that will enable rapid shape comparisons of library compounds against a target receptor (e.g., estrogen receptors) or against other compounds (e.g., known estrogenic compounds). The proposed implementation of *Shape Signatures* automatically incorporates information on the three-dimensional structure of chemicals, conformational flexibility, and incorporates charge-based information (including hydrogen bond donors and acceptors) into a unique signature for each chemical. The *Shape Signatures* methodology is outlined in Figure 1. *Shape Signatures* were applied in a number of typical EDC risk screening scenarios, involving both ligand- and receptor-based strategies. This new technology is fully applicable to many classes of EDCs such as androgens, retinoids and thyroid hormones. The *Shape Signatures* was extended to locate compounds with structural data that are similar to compounds of known estrogenic activity. Compounds will be identified that are complementary to the binding pocket within the ligand-binding domain (LBD) of estrogen receptors α and β. For this purpose, the available crystal-structure geometry of the ER-α LBD and a 3-D homology model of the ER-β LBD is being employed. The ultimate goal is to develop a tool for general use by scientists involved in risk screening and priority setting activities, including those who are not experts in computational chemistry or modeling. By its design, this tool is intended to be easy to use, fast, extensible, physically intuitive, and visually accessible through a graphics user interface (GUI).

Investigators

William J. Welsh - University of Missouri - St Louis

For More Information

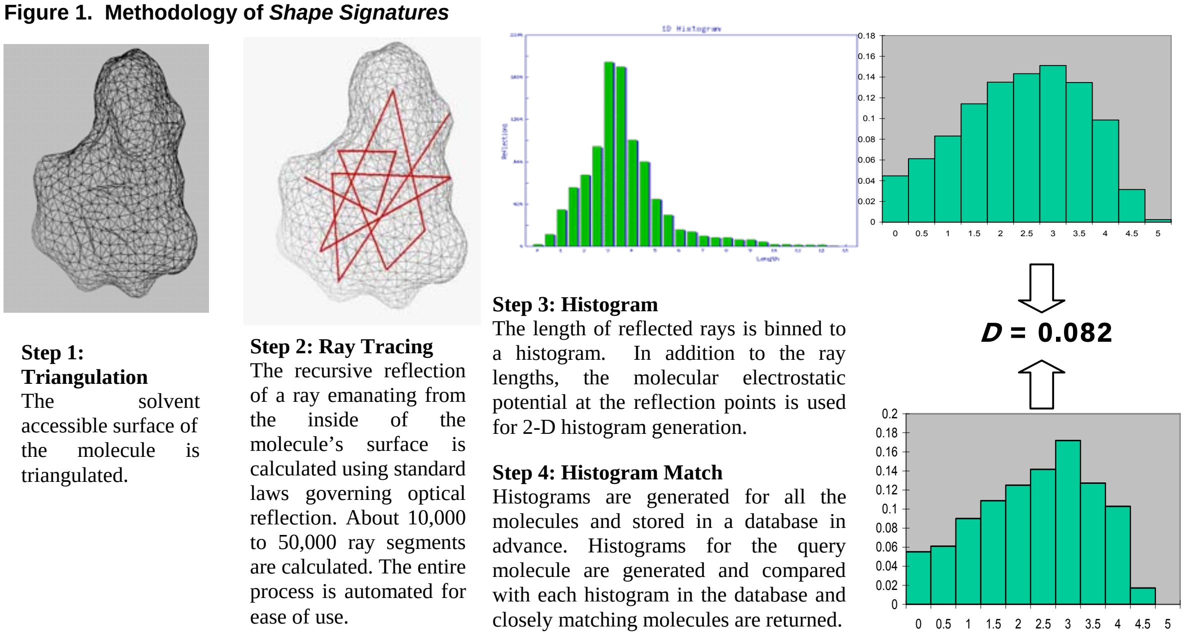
Laboratory web page:

http://www2.umdnj.edu/pharmweb/Faculty/jwelsh/jwelsh.htm

NCER Project Abstract and Reports:

http://cfpub2.epa.gov/ncer_abstracts/index.cfm/fuseaction/display.abstractDetail/abstract/169/report/0

Figure 1. Methodology of *Shape Signatures*



Step 1: Triangulation
The solvent accessible surface of the molecule is triangulated.

Step 2: Ray Tracing
The recursive reflection of a ray emanating from the inside of the molecule's surface is calculated using standard laws governing optical reflection. About 10,000 to 50,000 ray segments are calculated. The entire process is automated for ease of use.

Step 3: Histogram
The length of reflected rays is binned to a histogram. In addition to the ray lengths, the molecular electrostatic potential at the reflection points is used for 2-D histogram generation.

Step 4: Histogram Match
Histograms are generated for all the molecules and stored in a database in advance. Histograms for the query molecule are generated and compared with each histogram in the database and closely matching molecules are returned.